DOCUMENT RESUME

ED 189 099 TM 800 158

AUTHOR Forster, Fred
TITLE Everything You Wanted to Know about the Rasch [Model] (But Were Afraid to Ask). Portland Public Schools Occasional Papers in Measurement No. 17.
INSTITUTION Portland Public Schools, Oreg.
PUB DATE Jun 78
NOTE 11p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests: Educational Objectives: Elementary Secondary Education: Equated Scores: Item Analysis: Item Banks: *Latent Trait Theory: *Testing Problems: *Testing Programs: Test Interpretation
IDENTIFIERS Portland School District OR: *Rasch Model

ABSTRACT

Criticisms of norm referenced standardized tests are contrasted with the assumption and claims of proponents of Rasch model testing and item banking. The Rasch test theory model is based on the assumption of equal interval curriculum referenced scales for both students' abilities and test item difficulties. The Portland school district has established eight forty-item achievement tests, each covering only a short range of achievement, but all linked to a single ability scale. It is also possible to construct special purpose tests by selecting items from the item bank, since all items are keyed to achievement goals. Among the advantages of using the Rasch model are that you don't need representative samples for item analysis, and scores on the Rasch ability scale have tolerance values to indicate the expected size of measurement errors. The Portland school district has conducted a number of experiments with Rasch Statistics and concluded that samples of 200 cases are sufficient for item analysis, and to estimate the accuracy of the equating across ability levels. Work on additional curriculum areas and on rapid and meaningful score reporting is being developed. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED189099

PORTLAND PUBLIC SCHOOLS

OCCASIONAL PAPERS IN MEASUREMENT

OCCASIONAL PAPER NO. 17

EVERYTHING YOU WANTED TO KNOW ABOUT THE RASCH (BUT WERE AFRAID TO ASK)

Fred Forster
Portland Public Schools
June 1978

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Portland Public Schools
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM800 158

EVERYTHING YOU WANTED TO KNOW ABOUT THE RASCH

(but were afraid to ask)

Fred Forster
Portland Public Schools

It is no news to those of us who work in the schools that the rules of the testing game have changed drastically in the past decade. Political pressures at the state and national level, community pressure on our Boards, and increased visibility in the press have led to a variety of innovations designed to shore up our testing programs. Trying to make these innovations work, however, has uncovered weaknesses and limitations in the traditional methods we have trusted for years.

For example, several large school systems have tried out-of-level or ungraded testing designed to give each student a test matched to his achievement. Unlike some testing notions, this idea has as much basic appeal to the teachers as it does to the testing guru. From the teacher's viewpoint, each student is given a chance to feel successful with the test and has less reason for causing trouble during the test period. From the guru's viewpoint, students are able to achieve scores closer to the magic 50% level, so test reliability is improved. To make the program work, however, it is necessary to overcome a multitude of problems.

First, the program requires a foolproof way of getting the right test to the right student and the right code on the right answer sheet. This entails the usual array of data processing headaches inherent in getting the machine time and programs necessary to use last year's results to determine this year's test (i.e., lost kids, lost classes, program errors). In addition, since few publishers design their tests for this type of program, keeping track of the time each student spends on a test can become a nightmare. However, with a little lead time, the ingenuity of teachers and administrators usually comes to the rescue, and this obstacle is handled neatly.

The second step in launching the off-level program is devising a way to report class, school and district averages. Since several different tests may be used in the same classrooms, adding raw scores together produces a hodgepodge. On the other hand, the questionable practice of averaging grade equivalent scores for students across the various test forms can get you and your Superintendent in hot water. Since grade equivalents are extrapolations of group medians which gloss over the tremendous variance in performance at a grade they represent, at best, a rank order scale. Even worse, if a large number of students are shifted to a lower level test form (which almost always happens when your district has a sizeable number of inner-city students) the district averages drop dramatically because of the quirks built into the grade equivalent process. Needless to say, explaining that this drop is due to a statistical artifact is not likely to raise your popularity with either the Superintendent or Board.

A more statistically respectable approach to combining scores across test forms, which has been advocated by test publishers, is the use of the equating tables. The Anchor Study, a most ambitious and respectable use

traditional equating methods, has demonstrated the full array of difficulties and the tremendous costs involved in working with even a few tests in a single school subject. Looking at the money spent on the Anchor Study, remembering that it covered less than twenty tests at three grade levels, and considering that all those tests are now out of date, it boggles the mind to think of the cost for norming and equating tests across all subjects and grades. Beyond cost considerations, the fact is that traditional test equating methods just aren't very good. A quick check of the spread of scores at the top and bottom of the tables usually suffices to convince even diehard believers that traditional equating methods have serious limitations.

The point of the discussion to this point is simple; while testing innovations promise to benefit all, they usually are big disappointments because of the limitations or inflexibility of traditional test analysis procedures.

In our own district, we developed a laundry list of capabilities needed to bring flexibility to our testing program. The list included:

(1) a better way to specify the level of an item without having to randomly sample or overtest any group of students,
(2) a better way to develop parallel test forms to insure test security,
(3) a better way to report student achievement with reference to the curriculum of a subject rather than in relation to a hypothetical, non-existent reference group,
(4) a better way to tie norm referenced and criterion referenced test results together to eliminate criticisms about double-testing and over-testing, and
(5) a better way to report student progress to teachers in terms of overall achievement and achievement on specific learning goals.

The Rasch Model

Since traditional test analysis methods had disappointed us so often, we took a hard look at the Rasch testing model. Its most attractive feature was that it was based on the notion of establishing an equal interval curriculum referenced scale of student performance. It was claimed that this shift in perspective gave the Rasch a number of advantages over traditional approaches. First, the level of an item could be calibrated without random sampling. While traditional p-values vary widely from group to group, the Rasch item level calibration was purported to be constant from group to group and test to test. Similarly, while the same raw score on two different levels of student performance, it was claimed that the Rasch achievement level for a student was constant from test to test. This came about because the student's performance on the test was related to the learning continuum (i.e., his ability to read or make basic computations) not just his standing relative to other students.

Since the Rasch scale is anchored to the curriculum it seemed to provide the flexibility needed for item banking. Essentially, an item bank is a shift from the test as the unit of analysis to the item as the fundamental unit of analysis. While this may seem a modest change in

emphasis, it cannot be accomplished using group dependent methods.

After working with Rasch methodology for five years and testing its claims extensively, we have found that they are supported by the research evidence. Using previously calibrated items we have found it is possible to accurately predict how new tests will work even before they are tried with students. Similarly, we have found that the item bank has significantly reduced the time needed to construct a survey test or criterion test, while introducing a new level of precision to the selection of items. As a bonus since the tests are drawn from items calibrated to a single equal interval scale, they are automatically equated. The results of our research provides strong evidence that the learning-based/population-independent nature of Rasch item levels eliminates the need for elaborate sampling designs, and for traditional test equating methods.

The Practical Value of the Rasch Model

While the advantages of the Rasch approach have been discussed for more than a decade, we have just begun to learn how to use it to solve practical problems. Five years ago, the Northwest Evaluation Association began stockpiling items in reading and mathematics to build its item banks. Relying on the fact that measurement is most effective when a student can handle the test he receives, the Portland district has moved from the use of an eighty or one-hundred item survey test (where thirty or forty items are either too easy or too hard for the student) to a set of eight forty-item "level" tests on which items cover a short range of achievement. By assigning the tests to give each student items near his achievement level, we have been able to cut the testing time in half and still maintain a comparable level of reliability. As a fringe benefit, since the student doesn't encounter items that are overly hard or easy, he is less likely to guess wildly, become bored, or cause a disturbance.

In addition to guiding the construction of the level tests, the Rasch model was instrumental in the analysis and reporting of test information. Since all forms were drawn from the same item bank, they were tied to the same underlying equal interval scale. Once the student raw scores were translated into the new scale, they were averaged by class, school and district without further worry as the to test taken. This feature of the Rasch makes it possible to compare achievement from year to year even though students move on to higher level test forms.

In a similar way the Rasch model has provided us with a variety of advantages in the way we construct criterion referenced tests. Since goals are specified for each item in the bank, it is possible to select items which measure a goal at any specified achievement level. As an added bonus, since these tests are often drawn from the same bank as our norm-referenced tests, the criterion referenced results relate to those from the survey program to provide a continuous history for the student. Thus the Rasch simplifies the work needed to develop CRTS, and unifies the record keeping for our testing programs.

The Rasch also provides a solution to the problems of test security and of updating tests to reflect changes in district curriculum. By using the same blueprint to generate as many parallel forms as needed

(depending on the number of items available in the bank) test security is handled easily. Since the forms relate to the same learning scale, there is no additional equating work to be done. In revising a test form, we first field test items representing new curricular areas in which we are thin, and then modify the test blueprints to pickup the new goals and drop the outdated ones. Since both the old form and the revised form relate to the same underlying scale, the need for renorming is eliminated.

What is the Rasch Model

The Rasch approach is named after Georg Rasch, a Danish mathematician. In our own country, Benjamin Wright of the University of Chicago, has been the chief spokesman for the Rasch among the university folk. Ben has refined many of Rasch's original ideas and has organized several workshops designed to inspire interest. Primarily through his efforts, interest in the Rasch has grown among school test directors and measurement practitioners and it is being applied in a variety of educational settings.

The strength of the Rasch model lies in the assumption that a school ability such as reading can be measured on an equal interval scale in the same way height can be measured in feet or weight in pounds. This premise can be viewed as a generalization of the familiar notion of "true score" from traditional statistics. The advantage of Rasch's assumption is that he looks beyond this particular test, to a student's "true" achievement in reading or mathematics which has led to his performance on this test. A second assumption made by Rasch is that the success a student has with a given test item depends on his achievement level and the difficulty level of the item. Thus, we expect (and find) that the best students do well on all but the highest level items, and the lowest students do well on only the lowest level items. Besides the natural appeal of this assumption, it resembles the familiar notion of item discrimination traditionally used to judge the quality of items.

Using these simple ideas and some high powered mathematics, Rasch has transformed sample bound p-values and raw scores into statistically independent equal interval scaled measures of the level of an item and the achievement of a student. Thus, while item p-values vary from group to group, Rasch item levels are consistent even for groups with quite different achievement levels. From a practical point of view, this eliminates the need to secure random or representative samples when doing field testing, and opens the possibility of measuring a student's progress in terms of performance on the curriculum of a subject, rather than solely in terms of a national sample.

In one sense the Rasch model as a natural extension of traditional test theory from rank order measurement to equal interval measurement. The traditional concern with error of measurement is improved and expanded in terms of a unique "tolerance" value (measurement error) specified by the Rasch for each of the possible scores on the test. As you might expect, the tolerance is smallest for scores near the middle of the test and largest for scores which are very high or very low. The Rasch model provides a richer concept of test reliability. The traditional formulations of reliability are based on the quality of measurement provided by a test when given to a specific group of students, i.e., the reliability is as much a reflection of the group given the test as of the test being administered. Like other correlations between scores, therefore, the reliability plummets when the range of students score is small. This

limitation leads to the contradiction that as students are measured more and more accurately by assigning tests so that nearly every student is well placed, the traditional reliability will drop! For that reason, we now use an "absolute" reliability coefficient, based on the tolerance in the scores and operating range of the test, which always reflects the improved quality of our measurement

Traditionally, the correlations between performance on the item and performance on the test, (the point biserial) has been used to judge the effectiveness of an item (items with low point biserials are suspect since they don't pull their share of the measurement load). Like tradional reliability, however, the point biserial is affected adversely when the range of student scores is restricted or the percent correct on the item is very high or very low. The Rasch approach avoids these problems by substituting the notion of the degree of fit between how an item should work and how it actually works. In practice, the Rasch index of fit can be extremely useful in locating subtle defects in an item, such as an unintentional tipoff which helps low students get the item right for the wrong reason.

In our work we use traditional and Rasch statistics to evaluate the quality of items and the adequacy of tests, and use the item analysis for high and low student groups to complete the picture. Thus, rather than replacing methods we have used, the Rasch provides an added dimension, which makes it possible to do a better job.

Evidence for the Rasch Model

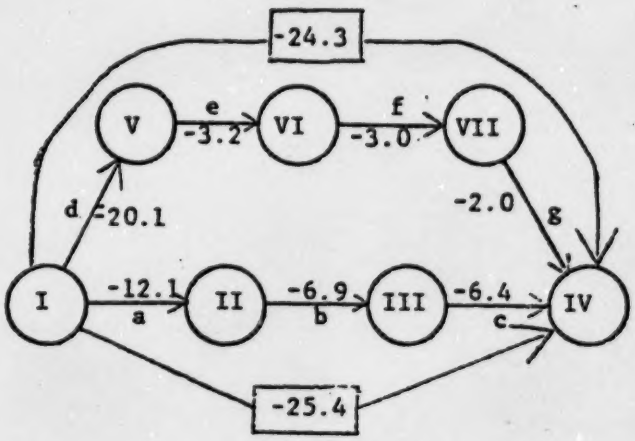
Before actually applying the Rasch method, we found it necessary to determine whether it really worked. Our first research project was to determine whether item calibrations were actually independent of the students taking the test. We split approximately 1400 fourth grade reading and eighth grade mathematics tests into two groupings: (1) students scoring above the district average vs. students scoring below the district average and (2) students in inner-city schools vs. students in our more advanced schools. The calibrations for each subgroup were compared to those for the total population, the resulting correlations were amazingly high (all larger the .95). Since groups as extreme as those in this experiment rarely occur in an acutal setting, it appears that Rasch item calibrations are actually independent of student characteristics in practical school situations. More recently we cross validated this experiment with 4000 students at grades four through eight in reading and mathematics have found the same results.

The second research project we undertook was the determination of the optimum sample size for calibrating items. Again we started with approximately 1400 students tests in fourth grade mathematics and eighth grade reading. We drew five random samples each of 300, 200, 100, and 50 from the total group and calculations for each sample. Our analysis of the correlations and standard deviations between the sample calibrations and population calibrations indicated that about 200 students were needed to obtain item calibrations equivalent to those for the total population. We cross validated this experiment last year at grades four and seven in reading and mathematics with approximately 4000 students at each grade and the results confirmed the 200 figure.

Our third research project was to verify that the calibrations for an item would remain constant regardless of the range of items on the test. To verify this we started with 1400 students in fourth grade mathematics We systematically whittled the original test down from 80 items to 60, 50, 40, 30, 20, and 15 items eliminating the items at both extremes As the range of the test decreased, the calibrations remained the same (standard deviation differing by less than .02%, and correlations in excess of .990), indicating that test range does not present a problem to Rasch scaling.

The fourth research project was to verify that the item calibrations remained constant from test to test. The pilot project involved creating a set of seven interlinking tests, i.e., each test had two blocks of items that appeared on the other tests.

TEST I	a	d	x
TEST II	a	b	x
TEST III	b	c	x
TEST IV	c	g	x
TEST V	d	e	x
TEST VI	e	f	x
TEST VII	f	g	x



x-indicates items included in test but not included in linking pattern.

As shown in Figure 1, the blocks of items shared between tests make it possible to evaluate whether the items retained the same calibrations across tests (and student groups). In addition, the links between tests could be cross validated (using two other links in the pattern) to assess consistency across the total linking network. The results from this experiment indicated that the linking values via one route (TEST I - TEST II - TEST III - TEST IV) were virtually identical to those via an entirely independent route (TEST I - TEST V -TEST VI - TEST VII - TEST IV). This notion of linking and cross checking links is not built into our field testing. After examining several test networks forty or more field tests each, we have found a similar level or accuracy and consistency.

The crucial confirmation of our research came with the demonstration that different tests provide the same achievement level for our students. We asked students in grades three to eight to take a forty-item field test in addition to the regular 80-to-100-item standardized reading test. This resulted in approximately 250 student pairings for each of two tests at each grade level. We then calculated the student's achievement level using only the field test and then using only the standardized test. While each student's raw scores were quite different, the mean achievement levels for the pairings were virtually identical. This research was replicated as part of the routine analysis of our district's new level test program. In this case there were fourteen levels of reading and nineteen levels of mathematics tests administered to approximately 4000 students at each grade level four through eight. The results showed near perfect agreement between the averages for these tests and a short grade level test in both of the subjects.

It may worth-while at this point to give a brief illustration of one of the ways the Rasch approach can refine a traditional concept. In an exploratory study, we isolated the most discriminating and least discriminating items on one of our standardized tests. When we checked out the traditional statistics for the two subtests, it was found that the traditional standard error of measurement was highest for the high discrimination subtest, i.e., our measurement appeared to be least accurate when the best items were used. The Rasch analysis, however, yielded a more reasonable conclusion. It indicated that the average tolerance for the student scores was lower for the high discriminating items, i.e., more accurate. This is one example of the way in which the Rasch approach has helped eliminate some of the paradoxes we have found in the use of traditional analyses.

To summarize, five years of research involving hundreds of tests in reading, mathematics and language arts and thousands of students continues to support the claims made for the Rasch model by Ben Wright and others.

Related Test Models

Some recent articles have drawn attention to a variety of "latent trait models" including the Rasch. These related models differ from the Rasch in that they add extra variables in the hope of better describing how an item works with students. Since the Rasch makes no provision for differences in item discrimination, and items obviously differ in this diminsion, we wondered whether including an extra variable for item discrimination might improve our results. Reviewing the available research we found two studies based on theoretical data which indicated that there was little or no loss in the precision of Rasch item calibrations due to fluctuation in the item discriminations. To try this out with real students and tests, an experiment was planned in which the items from a fourth grade mathematics test were separated into low, middle, and high discriminating categories so that each subgroupings satisfied the assumption of equal item discriminations. After calibrating each subtest separately, it was found that the item levels were the same as those obtained for the total test where the discriminations varied considerably. This research led to the conclusion that the differences in item discrimination encountered in practical situations have no significant effect on Rasch scaling.

Beyond the Rasch Model

The Rasch approach only one of several tools needed to develop to improve testing. Through the leadership of Drs. Doherty and Hathaway, The Portland district has been involved with the Tri-County Course Goal Committee in the development of collections of goals which span the full breadth of curriculum offered in a K-12 instructional program. Currently, collections have been developed for language arts (including reading), mathematics, science, social studies, home economics, art, music, industrial education, business education, second language, physical education and health education. Unlike most collections of behavioral objectives which tend to be uneven, the consistent level of definition maintained in the course goals makes the invaluable for defining the content of test items and for communicating with teachers. Taking advantage of the goals to specifiy the content of an item, completes the picture provided by the Rasch scaling, making the item bank a truly flexible tool.

A second area of opportunity within our testing programs is the development of better methods for reporting results to teachers, students, and parents. We have found that two kinds of test reports are needed by our teachers. The first is a report which covers the specific learning needs of each student; the second is a norm referenced survey report. Since the value the teacher derives from the report significantly influences the way in which the student approaches the testing task, we do not feel comfortable with the delays normally involved in scoring services available through test publishers. Our present district wide testing programs involves two stage reporting. We return a goal report to each teacher within a week of testing, following a week later by a report outlining the achievement level of the student and related norm referenced information. This approach is intended to get information to teachers when they need it to get students placed and settled, and begin working on the specific learning difficulties a group of students may have. A fringe benefit of this approach is the breather it provides our data processing department in handling the logistics of producing the reports and scheduling computer time.

A critically important part of our testing program is inservice. We have found that the gulf between the measurement course that teachers took in college and their classroom experience is so great that little appears to have rubbed off. The introduction of the new Rasch achievement scales has provided a useful excuse to work with teachers and help them feel more comfortable when interpreting test information. In the long run, since the combination of curriculum based and norm based reporting makes more sense to teachers we believe that our inservice is becoming more and more effective.

Summary

This discussion has meandered through some of the measurement ideas we have tried to implement in Portland District and the ways in which the Rasch approach has helped solve some practical measurement problems.

While the Rasch method represents a quantum leap forward in terms of our ability to measure and report student achievement, it is not a cure-all. Rather, it is the first step toward a more reasonable more practical, more comprehensive approach toward testing.